Preventing Racial Bias in Federal AI

Morgan Livingston
University of California Berkeley, Interdisciplinary Studies, Berkeley, California 94720
https://doi.org/10.38126/JSPG160205
Corresponding author: mlivingston@berkeley.edu
Keywords: AI; artificial intelligence; racial bias; diversity; impact assessment; contestability

Executive Summary: Artificial Intelligence (AI) systems are increasingly used by the US federal government to replace or support decision making. AI is a computer-based system trained to recognize patterns in data and to apply these patterns to form predictions about new data for a specific task. AI is often viewed as a neutral technological tool, bringing efficiency, objectivity and accuracy to administrative functions, citizen access to services, and regulatory enforcement. However, AI can also encode and amplify the biases of society. Choices on design, implementation, and use can embed existing racial inequalities into AI, leading to a racially biased AI system producing inaccurate predictions or to harmful consequences for racial groups. Racially discriminatory AI systems have already affected public systems such as criminal justice, healthcare, financial systems and housing. This memo addresses the primary causes for the development, deployment and use of racially biased AI systems and suggests three responses to ensure that federal agencies realize the benefits of AI and protect against racially disparate impact. There are three actions that federal agencies must take to prevent racial bias: 1) increase racial diversity in AI designers, 2) implement AI impact assessment, 3) establish procedures for staff to contest automated decisions. Each proposal addresses a different stage in the lifecycle of AI used by federal agencies and helps align US policy with the Organization for Economic Co-operation and Development (OECD) Principles on Artificial Intelligence.

I. What is racial bias in AI and why is it a problem?

Federal agencies are increasingly adopting Artificial Intelligence (AI) and delegating critical decisions to the technology. Out of the 142 largest federal agencies, 45% use or have considered using AI, for tasks ranging from setting bail to detecting fraud (Engstrom et al. 2020). Although AI can bring efficiency and objectivity to services, AI systems can also magnify systemic inequities. AI can replicate and amplify existing biases, producing predictions with harmful outcomes for a racial group. The causes for bias are both technical and social: the code can be embedded through the biases of the designers and data, and the use of AI can exacerbate bias already existing in a social system.

When used by a federal agency, AI predictions take on power as the basis for critical decisions, or in the case of automated decisions, the cause for immediate impact. The lifecycle of an AI system is an iterative process of defining the problem AI addresses, deciding to use AI, designing, coding, testing, deploying, maintaining and retiring the AI. The impact of an AI system depends on choices made at different stages in the AI lifecycle, including:

- Designers train an AI model to form predictions based on patterns learned in historical data, choosing the dataset the model will learn from, the accuracy of the model's prediction for different groups, and the testing procedure to evaluate the model.
- Staff deploys the AI system for their use case, choosing whether the AI model is appropriate for their task, how to use the AI predictions and who will manage the AI.
- Users act on the AI predictions, choosing how to manage the AI system and translate the machine output into conclusions with real impact.

Without sufficient safeguards, human choices can incorporate racial bias into AI systems, causing significant impact. Studies show racial bias in AI has already caused harm in many sectors including facial recognition, criminal sentencing, healthcare, and financial services.

- Facial Recognition: Facial recognition tools produce significantly higher false positive rates for African and East Asian individuals than for white individuals (Grother et al. 2019). One commercial tool had a 0.8% error rate for light-skinned males, but 34.7% error rate for dark skinned-females (Buolamwini & Gebru 2018). Disparate errors can lead to law enforcement falsely matching suspects with criminal databases (Snow 2018).
- Criminal Sentencing: COMPAS, an automated risk assessment tool used for criminal sentencing in Arizona, Colorado, Delaware, Kentucky, Louisiana, Oklahoma, Virginia, Washington and Wisconsin, incorrectly labeled black defendants as future criminals at close to twice the rate as white defendants (Angwin et al. 2019).
- Healthcare: A healthcare algorithm responsible for 200 million people systemically prevented almost 30% of eligible black patients from receiving additional care by giving lower risk scores to black patients than white patients with equal diagnoses (Obermeyer et al. 2019).
- Loans: FinTech firms charged Latinx and African-American loan borrowers 7.9 and 3.6 basis points, respectively, more than equivalent White borrowers, costing a yearly extra $765 million in interest (Bartlett et al. 2019).

The examples above are a selection of the known cases where biases in the design or use of AI led to racially disparate harm.

II. Causes

The societal impact of an AI system depends on the technical design, deployment for a specific task, and use by the staff. At each stage of the AI lifecycle, there is a risk that bias may be incorporated into the system.

i. Design

During development, the lack of racial diversity in technologists means the needs of the impacted populations are not represented in the design process (West 2019). Black and Hispanic workers are underrepresented in technology development, making up 8.1% and 5.8% of the national computer and math workforce respectively (National Science Board 2019). At Google only 2.8% of the technologists are Black (Google 2018), and at Microsoft only 3.3% are Black (Microsoft 2019). This lack of diversity hinders the foresight of teams responsible for creating tools that must integrate into complex and diverse social environments.

AI predictions are more accurate when the model is trained on large amounts of data. However, datasets often fail to represent the diversity of populations affected by the output. When a facial recognition system is trained on a set of photos that are primarily white and male, the system will be better at matching white male faces, and worse at matching black female faces (Buolamwini & Gebru 2018). Datasets can also reflect existing disparities in society, causing the model to learn pre-existing biases and output predictions exacerbating inequalities. For example, some predictive policing software uses data on past location of police responses to predict future locations of illegal activity. Police response data is an inaccurate predictor of illegal activity and often overrepresents communities of color. The algorithm flags historically over-policed areas for future police monitoring and creates a positive feedback loop reinforcing racially biased policing (American Civil Liberties Union et al. 2016). Identifying bias in datasets could mitigate development of biased AI, however there is no standard method to document the composition and source of datasets (Gebru et al. 2018).

ii. Deployment

The impact of an AI system is a product not only of design, but of its interaction with the dynamics and inequities of a specific social system (Crawford & Calo 2016). Without a social impact assessment incorporating domain expertise and community input prior to deployment, AI can produce unintended effects. A widely used healthcare algorithm systematically gave black patients lower scores because the designers used past healthcare costs to predict patients' future level of illness (Obermeyer et al. 2019). Black individuals historically have lower health care costs due to unequal access to care, and this inequality was replicated in the algorithm's prediction. If stakeholders and experts familiar with black communities' history with healthcare had been involved in assessing the algorithm, this bias may have been identified prior to the algorithm's deployment. In the absence of federal standards for AI fairness (Newman 2019), select companies have developed their own ethics and assessment guidelines for AI. However, these standards vary by company and, without enforcement, there is no guarantee of adherence (Hagendorff 2020).

iii. Use

When an agency incorporates an AI system into a decision-making process, power transfers from the staff to the classifications made by the AI code (Citron 2007). Some AI models are so complex that staff cannot understand how the AI functions and are unable to monitor the AI for error. If a deployed AI system outputs biased results, it can be impossible to retroactively determine the technical cause, either because of the model's complexity or because trade secret protections allow companies to hide proprietary code (State v. Loomis 2016). For example, Securities and Exchange Commission staff responsible for reviewing and acting on AI predictions for fraud detection do not always understand the reason for predictions due to the complexity of the AI. Customs and Border Patrol has been unable to determine the cause for error in its AI-enabled iris scanning tool because the code is proprietary to an external contractor (Engstrom et al. 2020). Even when staff review the AI prediction prior to any action being taken, the staff may overlook errors due to 'automation bias,' the human tendency to place too much trust in automated decisions in spite of contradicting evidence (Goddard, et al. 2012). However, research shows that providing documentation on the rules AI uses to make a prediction and training staff to understand the limitations and logic of the AI model increases vigilant monitoring and decreases risk of automation bias (Goddard, et al. 2012).

III. AI Standards Landscape

The US government has repeatedly called attention to the need to mitigate bias in AI, but has not followed up with sufficient action. In 2016, the White House report on "Preparing for the Future of Artificial Intelligence" highlighted bias as an issue and recommended creating representative datasets, examining diversity in the AI workforce, and verifying that the AI used by federal agencies is fair. Driven by the mandate of an executive order (US President 2019), the National Institute of Standards and Technology (NIST) issued a report recommending that federal agencies examine the use and impact of AI and adopt AI standards to minimize bias. The 2019 National Artificial Intelligence Research and Development Strategic Plan recommended federal agencies invest strategically to promote both contextual impact assessment of AI and record-keeping on AI development. The 2020 Office of Management and Budget Guidance for Regulation of Artificial Intelligence Applications directed agencies to mitigate bias through adopting tiered risk management for AI and adhering to standards (Vought 2020). NIST and other non-regulatory standards bodies are currently developing such standards to address bias, but there is little evidence federal agencies will be prepared to incorporate the guidelines. Despite these calls to mitigate bias in AI, government action has focused on lowering barriers to industry, not securing against bias. There is still limited to no federal policy ensuring equity for AI outcomes; the AI workforce lacks diversity; AI is neither tested nor documented sufficiently; and out of 64 of the largest federal agencies using or having considered using AI, none have established protocols to assess the potential impact of bias (Engstrom et al. 2020).

In 2019, the US government made an international commitment to protect against bias in AI and must now take action to follow through. The US, and over 40 other countries, agreed to the OECD Principles on Artificial Intelligence to ensure non-discrimination, equality, diversity, and fairness of AIs. The US endorsed increasing the inclusion of underrepresented populations, applying risk management to each phase of the AI lifecycle, investing in representative datasets, and enabling humans to challenge AI determinations (OECD 2019). Similar principles were then adopted by the G7 and G20 (G20 Trade Ministers and Digital Economy Ministers 2019). To uphold its commitment, the US government will need to make rapid progress in diversity, inclusion, AI assessment and practices for AI use.

In the absence of regulatory direction, companies and academia have worked toward ethical standards as a proxy for enforced policy (Calo 2017). Research on AI fairness has yielded mathematical formalisms (Dwork et al. 2012) and technical tools to expose the bias in models and datasets such as the IBM's AI Fairness 360 and the Google's What-If tool. However, fairness is ineffective as a broad standard because of its dependence on cultural, contextual and political values. AI fairness can be defined in as many as 21 different ways (Narayanan 2018), with different consequences for the design and outcome of an AI system. For example, achieving equal accuracy of a model for all groups may require decreasing the accuracy for certain groups (Barocas & Selbst 2016), creating a tradeoff with domain and case specific effects. Fairness must be determined according to the context of a specific AI use case, relying on domain expertise and stakeholder input on equitable outcomes.

IV. Policy Recommendations

Federal agencies must ensure the AI technologies used to assist or make decisions have equitable outcomes by preventing the racial bias that can arise in the design, deployment, and use of an AI system. As initial steps this memo recommends:

i. Federal diversity initiative for AI technologists

Preventing the development of biased AI from the outset requires increasing the racial diversity of the federal and federally-contracted technical development teams. To assess the true extent of the workforce diversity disparity for AI development, diversity reporting should be disaggregated by technologist role to expose the racial demographics of AI developers and researchers. Increasing diversity can be achieved by first requiring diversity reporting by role, to establish baseline statistics, then enforcing affirmative action for federal agencies and federally contracted employers engaging in AI development. Under the American AI Initiative federal agencies are prioritizing hiring for AI roles (White House Office of Science and Technology Policy 2020). Hiring inclusively is an opportunity for agencies to build diversity into the AI workforce of the future and build representation into AI development.

ii. Impact assessments for bias in AI

Before a federal agency implements or funds AI, they should conduct a standardized impact assessment to determine whether AI is appropriate for the use case and audit the AI system for bias. The depth of the assessment can be determined by the AI system's level of impact to individuals and communities (i.e. Appendix B of the Canadian Government's Directive on Automated Decision-Making; (Treasury Board of Canada 2019)). The assessment must include an audit of the data (Gebru et al. 2018), the programming, and any prior testing (European Commission 2020). The potential biases specific to the use case (Kim 2017) must be assessed by consulting the staff with experience on the social environment the AI will impact. Agencies should provide public notice and a period of public commentary to involve community stakeholders and to hold the agency accountable (Reisman et al. 2018). To ensure that the 33% of AI systems that are developed by private companies and then procured by an agency (Engstrom) are held to appropriate standards, the risk assessment can be implemented as part of a robust procurement procedure. Leveraging federal purchasing power in this way can shift the market (Calo 2017), promoting practices of record-keeping and thorough testing for private companies interested in federal contracts.

To support this depth of examination, agencies may need to incorporate increased technology expertise. Agencies will still need to incorporate assessments throughout the process of implementing AI, from deciding whether AI is the best use of resources, to crafting thorough Request For Proposals, reviewing potential changes to employee workflow, ensuring staff can maintain the technology and measuring its effect and costs over time (World Economic Forum 2019).

iii. Ensure contestability

Agencies must leverage the expertise of their staff to safeguard against harm from erroneous or biased AI predictions by ensuring staff can contest the AI prediction before action is taken (Almada 2019). Staff should be provided explanations of the rules the AI uses to make a determination, training involving hands-on experimentation to understand the function and limitations of the AI, and means to intervene or record errors in the prediction once the AI is in use (Hirsch et al. 2017). Continuous engagement enables staff to monitor for biased predictions and intervene when necessary, in line with the OECD principle of maintaining the 'capacity for human determination.' For more complex AI models whose rules cannot be easily described, further research is needed on providing practitioners an understanding of why a model acted a certain way in a specific context (Mittelstadt, et al. 2019).

V. Conclusion

The Use of AI is rapidly expanding and there is an urgent necessity for federal agencies to safeguard against its potential to effect racially disparate harm. With proper guidelines, AI can decrease agencies' costs, increase quality of services, and provide immense societal benefit. Increasing racial diversity in AI technologists, implementing AI impact assessments and enabling staff to contest AI decisions will help ensure AI systems produce equitable outcomes and enable federal agencies to realize the benefits of AI.

References

Almada, Marco. 2019. "Human Intervention in Automated Decision-Making: Toward the Construction of Contestable Systems." In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law - ICAIL '19*, 2–11. Montreal, QC, Canada: ACM Press. https://doi.org/10.1145/3322640.3326699

American Civil Liberties Union, et al. 2016. "Predictive Policing Today: A Shared Statement of Civil Rights Concerns." *American Civil Liberties Union*. August 31, 2016. https://www.aclu.org/other/statement-concern-about-predictive-policing-aclu-and-16-civil-rights-privacy-racial-justice

Angwin, Julia, Jeff Larson, Surya Mattu and Lauren Kirchner. 2016. "Machine Bias." *ProPublica*, May 23, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Barocas, Solon, and Andrew D. Selbst. 2016. "Big Data's Disparate Impact." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2477899

Bartlett, Robert, Adair Morse, Richard Stanton, and Nancy Wallace. 2019. "Consumer-Lending Discrimination in the FinTech Era." w25943. Cambridge, MA: National Bureau of Economic Research. https://doi.org/10.3386/w25943

Buolamwini, Joy, and Timnit Gebru. 2018. "Gender shades: Intersectional accuracy disparities in commercial gender classification." In *Conference on fairness, accountability and transparency*, pp. 77-91.

Calo, Ryan. 2017. "Artificial Intelligence Policy: A Primer and Roadmap," *U.C. Davis Law Review* 51, no. 2: 399-436. https://doi.org/10.2139/ssrn.3015350

Citron, Danielle Keats. 2007. "Technological due process." Wash. UL Rev. 85: 1249.

Crawford, Kate, and Ryan Calo. 2016. "There Is a Blind Spot in AI Research." *Nature* 538 (7625): 311–13. https://doi.org/10.1038/538311a

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. "Fairness through Awareness." In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference on - ITCS '12*, 214–26. Cambridge, Massachusetts: ACM Press. https://doi.org/10.1145/2090236.2090255

Engstrom, David, Daniel Ho, Catherine Sharkey, and Mariano-Florentino Cuellar. 2020. "Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies." *Stanford Law School*. https://www-cdn.law.stanford.edu/wp-content/uploads/2020/02/ACUS-AI-Report.pdf

European Commission. 2020. "White Paper on Artificial Intelligence – A European approach to excellence and trust." https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

G20 Trade Ministers and Digital Economy Ministers. 2019. "G20 Ministerial Statement on Trade and Digital Economy." https://www.mofa.go.jp/files/000486596.pdf

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. "Datasheets for Datasets." https://arxiv.org/abs/1803.09010v6

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). "Automation bias: a systematic review of frequency, effect mediators, and mitigators." *Journal of the American Medical Informatics Association: JAMIA, 19*(1), 121–127. https://doi.org/10.1136/amiajnl-2011-000089

Google. 2018."Google Diversity Annual Report 2018." https://static.googleusercontent.com/media/diversity.google/en//static/pdf/Google_Diversity_annual_report_2018.pdf

Google. 2020. "What-If Tool." *Google People + AI Research initiative (PAIR).* Accessed March 4, 2020. https://pair-code.github.io/what-if-tool/ai-fairness.html

Grother, Patrick, Mei Ngan, and Kayee Hanaoka. 2019. "Face Recognition Vendor Test Part 3::

Demographic Effects." NIST IR 8280. Gaithersburg, MD: National Institute of Standards and Technology. https://doi.org/10.6028/NIST.IR.8280

Hagendorff, Thilo. 2020. "The Ethics of AI Ethics: An Evaluation of Guidelines." *Minds and Machines*. https://doi.org/10.1007/s11023-020-09517-8

Hirsch, Tad et al. 2017. "Designing Contestability: Interaction Design, Machine Learning, and Mental Health." *DIS. Designing Interactive Systems (Conference)* vol. 2017: 95-99. doi:10.1145/3064663.3064703

IBM. 2020. "AI Fairness 360 Open Source Toolkit." *IBM Research Trusted AI*. Accessed March 4, 2020. https://aif360.mybluemix.net/

Kim, Pauline. 2017. "Auditing Algorithms for Discrimination." *University of Pennsylvania Law Review Online* 166: 189-204.

Microsoft. 2019. "Diversity and Inclusion Report 2019." https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE4aqv1

Mittelstadt, Brent, Chris Russell, and Sandra Wachter. 2019. "Explaining Explanations in AI." In *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '*19, 279–88. Atlanta, GA, USA: ACM Press. https://doi.org/10.1145/3287560.3287574

Narayanan, Arvind. 2018. "Tutorial: 21 definitions of fairness and their politics." *Conference on Fairness, Accountability, and Transparency*, NYC, Feb 23, 2018. https://www.youtube.com/watch?v=jIXIuYdnyyk

National Institute of Standards and Technology (NIST). 2019. "U.S. Leadership in AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools." https://www.nist.gov/system/files/documents/2019/08/10/ai_standards_fedengagement_plan_9aug2019.pdf

National Science Board. 2019. "Science and Engineering Labor Force." *National Center for Science and Engineering Statistics*. September 26, 2019. https://ncses.nsf.gov/pubs/nsb20198/assets/nsb20198.pdf

Newman, Jessica. 2019. "Toward AI Security: Global Aspirations for a More Resilient Future." *Center for Long-Term Cybersecurity*. https://cltc.berkeley.edu/wp-content/uploads/2019/02/CLTC_Cussins_Toward_AI_Security.pdf

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." *Science* 366(6464): 447–53. https://doi.org/10.1126/science.aax2342

Organisation for Economic Co-operation and Development (OECD). 2019. "Recommendation of the Council on Artificial Intelligence." Legal Instrument 0449. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Reisman, Dillon, Jason Schultz, Kate Crawford, and Meredith Whittaker. 2018. "Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability." *AI Now*. https://ainowinstitute.org/aiareport2018.pdf

Select Committee on Artificial Intelligence. 2019. "The National Artificial Intelligence Research and Development Strategic Plan: 2019 Update," *National Science and Technology Council*. https://www.nitrd.gov/pubs/National-AI-RD-Strategy-2019.pdf

Snow, Jacob. 2018. "Amazon's Face Recognition Falsely Matched 28 Members of Congress with Mugshots." *American Civil Liberties Union.* https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28

State v. Loomis, 881 N.W.2d 749, 765 (Wis. 2016).

Treasury Board of Canada. 2019. "Directive on Automated Decision-Making." https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592

U.S. Executive Office of the President. 2016. "Preparing for the Future of Artificial Intelligence," National Science and Technology Council, Washington D.C., October 2016. https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf

U.S. President. Executive Order. "Executive Order 13859 of February 14, 2019, Maintaining American Leadership in Artificial Intelligence." *Code of Federal Regulations*, 84(31), p.3967. whitehouse.gov/presidential-actions/executive-order-maintaining-american-leadership-artificial-intelligence/

Vought, Russel. 2020. "Guidance for the Regulation of Artificial Intelligence Applications." Executive Office of the President, Office of Management and Budget. January 2020. https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf

West, S.M., M. Whittaker, and K. Crawford. 2019. "Discriminating Systems: Gender, Race and Power in AI," *AI Now Institute*. https://ainowinstitute.org/discriminatingsystems.html

White House Office of Science and Technology Policy. 2020. "American Artificial Intelligence Initiative: Year One Annual Report." February, 2020. https://www.whitehouse.gov/wp-content/uploads/2020/02/American-AI-Initiative-One-Year-Annual-Report.pdf

World Economic Forum. 2019. "Guidelines for AI Procurement." http://www3.weforum.org/docs/WEF_Guidelines_for_AI_Procurement.pdf

Morgan Livingston is an undergraduate at UC Berkeley studying Technology Policy in the Interdisciplinary Studies Field and Data Science. Morgan is a research assistant at the Berkeley Center for Globalization and Information Technology and focuses on privacy and data law.